HYDROLOGICAL PROCESSES
Hydrol. Process. **21**, 1968–1976 (2007)
Published online 19 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.6489

Shared near neighbours neural network model: a debris flow warning system

Fi-John Chang,* Kuo-Yuan Tseng and Paulo Chaves
Department of Bioenvironmental Systems Engineering, National Taiwan University, Taipei, Taiwan, ROC

Abstract:

The main purpose of this study is to develop a new type of artificial neural network based model for constructing a debris flow warning system. The Chen-Eu-Lan river basin, which is located in Central Taiwan, is assigned as the study area. The creek is one of the most well-known debris flow areas where several damaging debris flows have been reported in the last two decades. The hydrological and geological data, which might have great influence on the occurrence of debris flows, are first collected and analysed, then, the shared near neighbours neural network (SNN + NN) is presented to construct the debris flow warning system for the watershed. SNN is an unsupervised learning method that has the advantage of dealing with non-globular clusters, besides presenting computational efficiency. By using SNN, the compiled hydro-geological data set can easily and meaningfully be clustered into several categories. These categories can then be identified as 'occurrence' or 'no-occurrence' of debris flows. To improve the effectiveness of the debris flow warning system, a neural network framework is designed to connect all the clusters produced by the SNN method, whereas the connected weights of the network are adjusted through a supervised learning method. This framework is used and its applicability and practicability for debris flow warning are investigated. The results demonstrate that the proposed SNN + NN model is an efficient and accurate tool for the development of a debris flow warning system. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS debris flow; warning system; shared near neighbours; artificial neural network; unsupervised learning

Received 30 December 2005; Accepted 16 June 2006

INTRODUCTION

Taiwan is located over the junction of the Eurasian and Philippine plates, having its accidental topography originating from the pressure of these two tectonic plates and resulting in high elevations and very steep slopes. As a consequence of such topography combined with frequent earthquakes in the region and regular typhoon seasons with constant torrential rains, Taiwan suffers from frequent floods, landslides, and debris flow events. Such events are frequently responsible for great economic damage, losses of human lives, and negative environmental impacts. Among all these catastrophic events, debris flows are surely among the most dangerous and devastating, as they combine the destructive characteristics of both the far-reaching impact of high tides and the powerful destructive momentum of landslides.

There have been many examples of catastrophes related to debris flows in Taiwan in the past. Unfortunately, the last few decades have witnessed an increase in the magnitude and frequency of these catastrophes. This is probably due to the rapid economic development and high population growth of Taiwan, which results in overexploitation of the environment including increased deforestation and occupation of flood plains and hillside areas (Yu, 2002). Basically, in every typhoon season or during extreme torrential rains, it is possible to observe the occurrence of debris flow events throughout Taiwan. In September 1989, for example, Typhoon Jasmine hit the central area of Taiwan, resulting in a death toll of 40 people and substantial economic losses. Another case was Typhoon Peach, which struck Taiwan in July 2001; it also caused tremendous damage and losses throughout Taiwan. As a result, more attention has been given to the study and prevention of such overwhelming catastrophes.

Debris flow can be defined as a mix of water with various solid particles such as sand, stones, gravel, rock stratum and rocks after the collapse of land and hillsides. The dynamics of the debris flow is quite different from water flows or the pure landslide phenomenon, behaving something like a mix of the two. It presents high velocities and has a sudden and powerful impact, which can often cause enormous destruction and damage. In general, the research in this field tends to investigate the basic causes of such flows, such as the geomorphologic formations, and to monitor and study ways and means for preventing and remedying such disasters through the use of proper technologies. Many of these researches focus on the application of topographical models, which aim to understand the basic mechanisms and physically influent factors. The physical-based model can be used to estimate not only the occurrence of the debris flow but the propagation mechanism as well. The approaches taken to predict landslide-producing storms may be summarized as: i) empirical analysis of such storms ii) empirical

* Correspondence to: Fi-John Chang, Department of Bioenvironmental Systems Engineering and Hydrotech Research Institute, National Taiwan University, Taipei, Taiwan, ROC. E-mail: changfj@ntu.edu.tw

mapping of landslide location and iii) mechanistic modeling of slope stability (Casadel *et al*., 2003).

Although empirical models may be simpler to develop and apply, their empirical nature may limit their general application. Crozier (1999) tested an empirical model for rainfall-triggered landslide prediction. The model intends to predict a 24-h landslide occurrence with a probability that the event will occur somewhere within the city within the following 24 h. Besides the empirical based approaches, there have been several works proposing the mapping of spatial characteristics of landslide potential, based on previously observed occurrences. These methods intend to produce static hazard maps. Usually, they are based on statistical distributed analysis considering previously observed events and assuming general conditions. To be able to handle such spatially distributed information, these approaches usually take advantage of geographical information systems (GIS) for extracting new information and generating maps by the overlaying of several other maps of different characteristics, such as topographic, land use, soil type, hydrological, and geological. Some examples of such an approach can be found in Chau *et al*. (2004); Lan *et al*. (2004); Perotto-Baldiviezo *et al*. (2004) and Ayalew and Yamagishi (2005).

Lollino *et al*. (2002) analysed landslide behaviour through cross-correlation methods between soil movement observation and rainfall events by using a complex monitoring network. They concluded that by observing the movements of the sliding surface, a high correlation with rainfall is revealed, identifying a time of 8–9 days between the occurrence of a rainfall peak and the corresponding peak in movement produced by this rainfall event. Casadel *et al*. (2003) applied a conceptual model similar to the TOPMODEL to an infinite slope to predict the spatial distribution of shallow landslides. And daily rainfall was also used to drive the model. They found that such models can perform better than simple threshold approaches. However, it requires considerable calibration efforts in finding appropriate physical parameters and deep knowledge about the processes involved. Approaches such as these are usually extremely costly, time consuming and depend on extensive data collection.

Landslides together with torrential flood waters constitute the main ingredients of debris flows. Debris flows are influenced by a range of factors including soil type, geomorphologic factors, land use, rainfall intensity and distribution, and topographic characteristics. Hence, the physical phenomena behind debris flows present great complexity and high non-linearity, making its accurate prediction an almost impossible task by only using traditional techniques. This is even more important knowing that the prediction of debris flows includes different sciences (e.g. hydrology, meteorology, surveying, geomorphology, and geology) and such flows occur within enormous spatial and temporal variability. Moreover, collection of related data, problems with extrapolation of local application and fully understanding all natural processes involved may result in an extremely costly and time consuming task.

One alternative to dealing with this problem of debris flow prediction may be found in the application of soft-computing techniques such as fuzzy theory and artificial neural networks (ANN). For instance, ANN is a very powerful tool in dealing with non-linear systems. It can mimic human learning and thinking capabilities through combining different sorts of input information to yield the desired output. In this research, we use some of the latest available technologies in the field of soft-computing for handling the complexities associated with the prediction of occurrence of debris flows in the development of an early warning system. So far, there has been very limited research about the prediction of landslide and debris flows using ANN-based models, and the existing one could only be found in very recent years. Some of the few successful examples of landslide issues in technical literature can be found in Lee *et al*. (2004); Ermini *et al*. (2005); Yesilnacar and Topal (2005) and Gómez and Kavzoglu (2005). Therefore, it is also important to expand and to continue the investigations on the applicability of these technologies in dealing with debris flows. A wider goal of the program was to harness better the mechanics and extent of the knowledge of the island debris flows. A specific goal of the study was to see the physical evidence of the island debris flows as a first step toward estimating risk of recurrence in the region.

METHODOLOGY

The constant improvement in computational efficiency of personal computers has brought about substantial and important development in science and technology, particularly in the areas of artificial intelligence or soft-computing. Many of these techniques are based on the concepts found in natural processes. For example, a genetic algorithm is based on the law of evolution, and swarm optimization is based on the capability of animals to find the 'optimal' path between their nest and the food source. Similarly, another type, if not the most popular one, involves ANN based on human brain structure. They are composed in neuron units, which take and process input information through their simple connections. The calculated results are transferred to the external world or other neural units by means of simple mathematical operations defined as transfer functions.

Generally, the most popular structure of an ANN-based model presents three layers, as illustrated in Figure 1, which comprise an input, hidden, and output layers. Usually, the input layer presents the same number of units as the available and identified variables in the input data set. The output layer will have as many units as the number of desired output variables. One of the advantages of ANNs is their adaptive nature in dealing with non-linear problems, such as highly complicated hydrological systems that are difficult to formulate and solve (Shamseldin, 1997; Chang and Hwang, 1999 and Sajikumar and Thandaveswara, 1999; Chang and Chang, 2001; Chang *et al*., 2005; Yang and Chang, 2005).

DOI: 10.1002/hyp

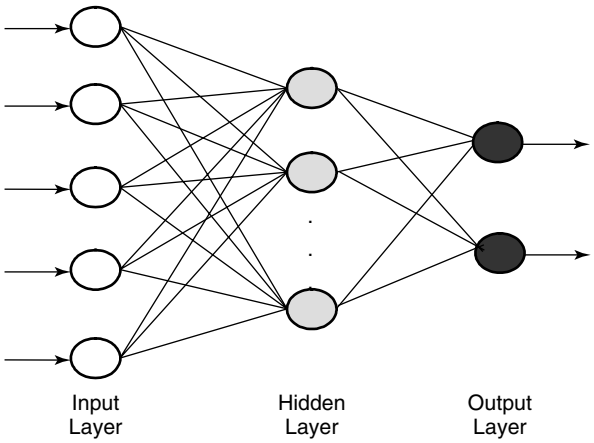

Figure 1. Example of a three-layer ANN architecture

In this research, we propose a novel type of ANN structure named the shared near neighbours neural network model (SNN + NN) and investigate its applicability in debris flow prediction. The SNN + NN model presents an input layer, a hidden layer based on a clustering method, and the output layer based on ANN. This structure is quite similar to the counterpropagation neural network (CPN), first proposed by Hecht-Nielsen (1987) and applied for stream flow reconstructing by Chang *et al.* (2001). However, differently from previous works, for the purpose of clustering, we have introduced the shared near neighbours (SNN) method, which was first proposed by Jarvis and Patrick (1973) based on an unsupervised training method that can be applied for large sample sizes with high dimensionality and non-globular clusters.

The proposed SNN + NN model can be separated into two stages, where input data are clustered to construct the rule bases during the first stage, and the second stage is responsible for training of the weights between the hidden layer units (clusters) and output units. One of the advantages of the SNN + NN model is that the clusters (or hidden layer nodes) required for building the model, can be generated automatically during the training process based only on two parameters. Thus, the number of nodes for hidden layers is actual. Obviously, the SNN + NN can be also recognized as a non-linear selforganizing model.

The shared near neighbours

The main characteristics of the SNN are as follows: i) it can automatically identify the number of clusters ii) it always presents at least one point of the data set in each cluster iii) clusters do not overlap and iv) if two points are sufficiently similar, then the probability that they come from the same cluster approaches unity.

Let (X_1, X_2, ..., X_t) be a set of data vectors in a p dimensional Euclidean vector space, and it is required that these t points be clustered into N groups. Data points are similar to the extent that they share the same near neighbours, which can be expressed by the parameter K. Hence, K specifies how many neighbour points to be considered when counting the number of mutually shared neighbours with another data point. K should be at least 2 and as a general rule, smaller values of K result in faster calculation but larger number of clusters. On the other hand, higher values may result in fewer clusters that form longer chains, but calculation is slower. The other important parameter of the algorithm is represented by KT, which can be referred to as the neighbours in common or similar threshold. This parameter specifies the minimum number of mutual nearest neighbours that should be present for two tested points to be considered belonging to the same cluster. KT should be at least 1 and should be smaller than K. K and KT are usually found through a trial-and-error method. To give an impression regarding the value and effect of K and KT, four clustering results of a two-dimensional butterfly profile are shown in Figure 2.

The cluster algorithm can be carried out in the following manner:

Step 1: For each point of the data set (X_1, X_2, ..., X_t), list the k nearest neighbours by order number (1, 2,..., k) according a measuring such as the Euclidean distance as shown in equation (1):

$$d_{ij} = \left\{ \sum_{l=1}^{p} |x_{il} - x_{jl}|^m \right\}^{1/m} \tag{1}$$

Where d_{ij} is the Euclidean distance between points i and $j(i, j = 1, 2, \ldots, t)$, t is the number of points in the data set, p is the dimensionality of the points, and m is the order number of the Euclidean distance. In this work, m is set equal to 2.

Step 2: Review the k nearest similar neighbours between two tested neighbourhood points. If both points have at least KT similar neighbours points, then they belong to the same cluster.

Repeat steps 1 and 2 until all the data points are clustered. The whole process would be carried out

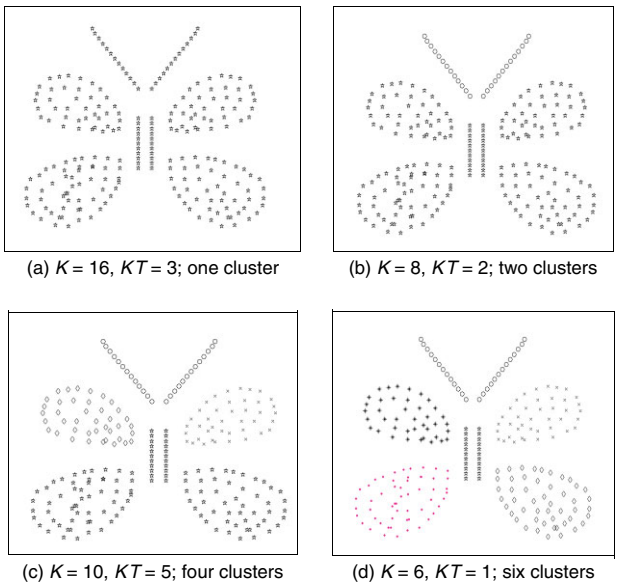

Figure 2. SNN Cluster results of a two-dimensional butterfly profile by using different values of K and KT

DOI: 10.1002/hyp

with new K and KT. Moreover, the comparison of the clustered results with actual observed information can be useful for defining the most appropriate ("optimal") values of K and KT. The SNN is then used to cluster the similar inputs into the same unit.

The SNN + NN procedure

Each cluster found at the previous section represents a hidden unit of the SNN + NN network and can be seen as a type of rule. The rule is picked by the comparison between the existing N clusters units and the input X(i), where i is the ith seeded inputs. The criteria of similarity are the closest distances between X(i) and the center of the j-th cluster w^j. If the closest distance between X(i) and w is smaller than 1, the X(i) is grouped to the j-th cluster unit. The algorithm for updated weight vector p is given by

$$\begin{aligned} &1.D(w^j, X(i)) = \min_{j=1\sim n} D(w^j, X(i)) \\ &2.if \quad D(w^j, X(i)) \le \Delta \\ &then \quad q^j_{new} = q^j_{old} + \beta[y(i) - q^j_{old}] \\ &else \quad q^j_{new} = q^j_{old} \end{aligned} \tag{2}$$

where D(w^j;$X(i)$) denotes the distance (such as the Euclidean metric shown in equation (1)) between the center of the j-th cluster unit w^j and the input $X(i)$, q is the weights between cluster j and the corresponded output unit. β is a positive constant update rate. Indices *new* and *old* represent the new calculated weights and the previous calculated weights q, respectively. The weights, q, are updated to reduce the error between the outputs of the network and the corresponding target outputs. $y(i)$ is the ith target value. Note that the first set of q weights is randomly generated.

After the structure of the SNN + NN is built, the model is used to calculate the outputs of the network. The predicting algorithm includes two steps. The first step is the pattern matching, while the second step determines the weighted average. Step 1 evaluates the similarity of X(i) and the rules. The similarity can be represented by:

$$s^j = S[X(i), (w^j, \Delta^j)] \tag{3}$$

where s^j is the similarity, $s^j \varepsilon$ [0; 1];. (w^j, Δ^j) is the jth rule with center w^j and interval Δ^j, and S denotes the similarity measure, which may be defined by:

$$s^j = 1 - D^j[X(i), (w^j, \Delta^j)] \tag{4}$$

where D^j ranges between 0 and 1 and is the relative distance from $X(i)$ to the j-th rule and is given by

$$D^j = \begin{cases} d^j/\Delta & if : d^j \le \Delta \\ 1 & otherwise \end{cases} \tag{5}$$

The Euclidean distance denoted by d^j, which is the distance between $X(i)$ and w^j, is defined by

$$d^j = \left[\sum_{l=1}^{p} (w^j_l - X_l(i))^2\right]^{12} \tag{6}$$

in which p is the dimensionality of the point. If d^j is greater than Δ, then set s^j to be zero. The rule j does not match the ith input and will not influence its output. An illustration of the above explanation, showing a graphic representation of the mentioned parameters and variables, is presented in Figure 3.

Lastly, the final output $Y(t)$ is deduced by weighted averaging, defined as

$$Y(i) = \frac{\sum_{j=1}^{N} s^j q^j}{\sum_{j=1}^{N} s^j} \tag{7}$$

SIMPLE TEST FOR THE SNN + NN MODEL

To test the efficiency and applicability of the proposed SNN + NN model a simple example is proposed. The example is based on the following function:

$$z = \cos(x) - 0.3 \cdot \sin(y) \tag{8}$$

for which a 450 points data set is built from randomly generated values of x and y from the 0 and 1 interval. For better evaluation of the model performance, the data set is divided into three groups; training, validation, and testing, having 300, 100 and 50 data points, respectively.

The first part of the proposed model, the SNN component, could well cluster the input information (x,y) into a number of clusters equal to approximately 1/10-th of the size of the training data (i.e. 32 clusters). This was found after trying different values of parameters K and KT in both training and validation groups. After a trial-and-error procedure the values of $K = 10$ and $KT = 6$ were believed to be the optimal values for the first part of the model. Figures 4–6 show a correlation plot of the generated values by equation (1) vs simulated values by SNN + NN for all three periods, respectively. The results

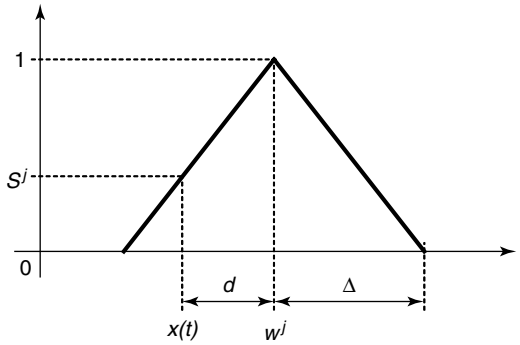

Figure 3. Graphical representation of the degree of similarity

DOI: 10.1002/hyp

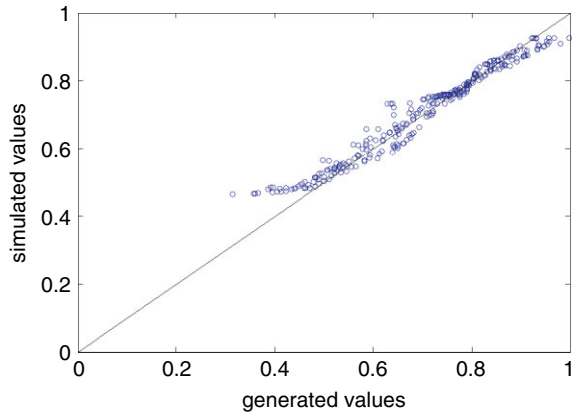

Figure 4. Correlation results between generated and simulated values–Training period

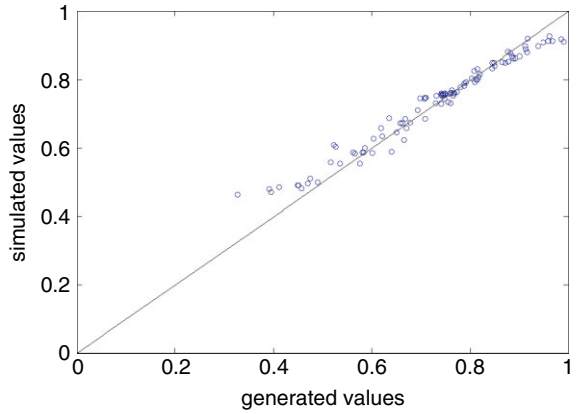

Figure 5. Correlation results between generated and simulated values–Validation period

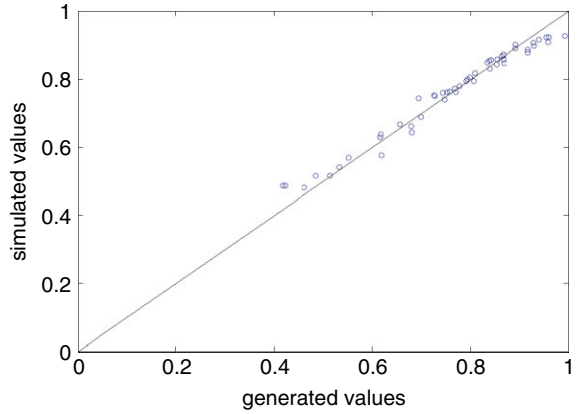

Figure 6. Correlation results between generated and simulated values–Testing period

have shown high accuracy in proving its ability to simulate a non-linear system. More than analyzing individual simulated and calculated points, focus should be given to the overall trends and generalization capabilities of the model, which apparently has been successfully demonstrated.

APPLICATION FOR DEBRIS FLOW PREDICTION

Study area

The Cheu-Eu-Lan river basin was used as the study area for further investigating and testing the developed SNN + NN model on the actual occurrence of debris flows events. The Cheu-Eu-Lan river basin is located in the Nan-Tou county in central Taiwan (as shown in Figure 7) and is in the high mountainous areas, with a height from 310 m to 2900 m with average height of 1500 m. The total length of the Cheu-Eu-Lan river is approximately 42.4 km with a drainage area of about 449.67 sq. km and average slope on the order of 1/20, indicating an extremely high steepness which results in a deep and narrow river with many torrential and rapid-flow tributaries. The topography and high precipitation rates have already been identified by previous studies as the main causes of debris flows in the regions.

Data situation and analysis

For practical application of the proposed model, data of observed debris flow events and potential influent factors have been collected. These data refer to 13 heavy rainfall events that occurred between 1985 and 1998. The data referent to the influent factors were then divided into two basic categories: hydrological and morphological, including i) effective rainfall duration and accumulated effective rainfall, and ii) area, length, slope and Horton's form factor, respectively. The morphologic data have been abstracted from a digital terrain model (DTM) topographic map (1 : 25 000 scale), using a geographic information system (GIS) tool for the 16 sub-basins

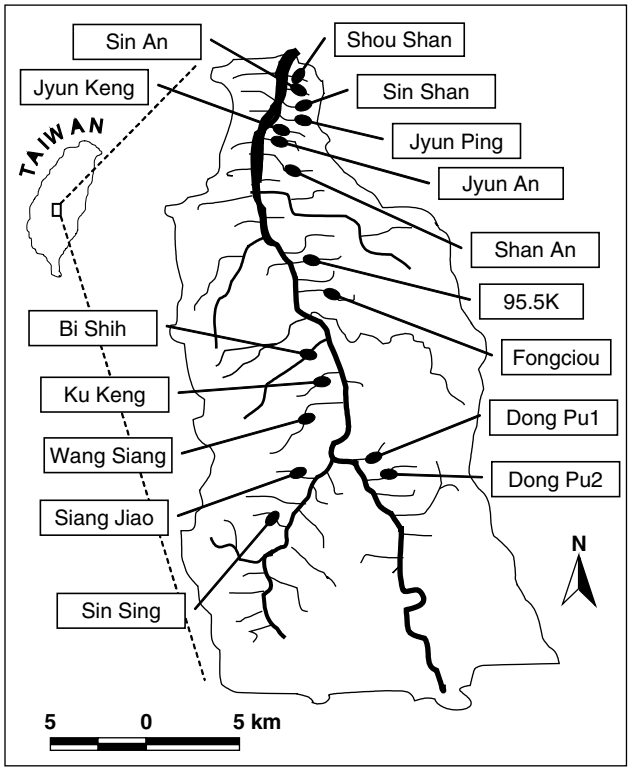

Figure 7. Location of Chen-Eu-Lan river basin and considered sub-basins

DOI: 10.1002/hyp

Table I. List of sites of debris flow and dates for available data

Location	Date and time	Rainfall event
Fongciou	23.8.1985–7 p.m.	Typhoon Nelson
Fongciou	22.8.1985–9 a.m.	Typhoon Wayne
Fongciou	31.7.1996–11a.m.	Typhoon Herb
Siang Jiao bridge	1.8.1996–1a.m.	Typhoon Herb
Sin Sing bridge	1.8.1996–1a.m.	Typhoon Herb
95.5K	1.8.1996 −1∼2a.m.	Typhoon Herb
Sin An bridge	1.8.1996 −2∼3a.m.	Typhoon Herb
Sin Shan bridge	1.8.1996 −2∼3a.m.	Typhoon Herb
Jyun Ping bridge	1.8.1996 −2∼3a.m.	Typhoon Herb
Jyun Keng bridge	1.8.1996 −2∼3a.m.	Typhoon Herb
Shan An bridge	1996.8.1 −2∼3a.m.	Typhoon Herb
Dong Pu 1 bridge	1.8.1996 −2∼3a.m.	Typhoon Herb
Fongciou	9.6.1998–6p.m.	Rain storm

Table II. Sub-basins, morphologic variables, and factors

Sub-basin name	Area (km^2)	Length (km)	Slope (degrees)	(Horton's) form factor
Sin shan bridge	0.54	0.67	24.7	1.21
Sin an	1.39	1.5	21.3	0.61
Jyun ping	0.88	1.26	23.3	0.55
Jyun keng	1.73	1.46	18.8	0.81
95.5K	1.64	2.03	24.7	0.4
Shan an	2.47	2.68	17.2	0.34
Fongciou	1.88	1.7	27.9	0.65
Siang jiao	2.2	2.1	22.3	0.5
Sin sing	2.18	2.79	23.3	0.28
Jyun an	1.07	2.38	20.3	0.19
Shou shan	0.64	0.82	21.3	0.95
Bi shih	1.44	1.59	11.3	0.57
Ku keng	1.24	0.87	11.9	1.64
Wang siang	1.59	1.99	12.9	0.4
Dong pu 1	0.65	0.96	23.3	0.71
Dong pu 2	2.54	2.52	26.6	0.4

where debris data are available (Table I gives a summary of the events, and Figure 7 the location of the sub-basins).

Available data on debris flow. Usually, data related to debris flow events are quite limited. Such flows tend to occur during high intensity rainfall periods, which make it very dangerous for onsite measurements. Moreover, they usually happen only on mountainous and less populated areas where observation stations are nonexistent. Table I shows the locations and dates for which extreme rainfall events have taken place and debris flows have been observed. With the data organized chronologically, it can be seen that the majority of the observed events occurred in the most recent years. This is probably due to the high and progressive development of hillside areas, which increases the chances of soil erosion, landslides and more rapid water flows. Therefore, the higher frequency of debris flow events in the past decades has also become of great concern and gained the attention of the general public.

Analysis of hydrological factors. Owing to the great uncertainties related to the timing and distribution of rainfall, different places usually experience singular effects caused by the same rainfall event. To properly consider the influence of a rainfall event that has been recorded at a certain rain gauge, different locations within the watershed are weighted by the distances between the sub-basin and the observation gauges. The method is based on the inverse of the square distance between the observation gauges and the point (sub-basin) of interest. Hence, the method can be described as follows:

(i) identify the coordinates x and y for the two parts (gauges and sub-basin) and calculate their distance d using equation (9), in which indices o and i represent observation and sub-basin points, respectively:

$$d_{oi} = \sqrt{(x_o - x_i)^2 + (y_o - y_i)^2} \tag{9}$$

(ii) define the inverse square ratio f as shown in (10):

$$f(d_{oi}) = 1/d_{oi}^2 \tag{10}$$

(iii) then calculate the weight λ to be applied to compensate for the distance between the observation gauge o and sub-basin i for all nearest gauge n, as represented by (11):

$$\lambda_i = \frac{f(d_{oi})}{\sum_{i=1}^{n} f(d_{oi})} \tag{11}$$

(iv) finally, calculate the actual compensated rainfall value Z' in sub-basin i after observed rainfall Z by using equation (12):

$$Z'(X_i) = \sum_{i=1}^{n} \lambda_i Z(X_i) \tag{12}$$

Rainfall intensity does influence the occurrence of debris flows. Nevertheless, it is possible to identify that effective rainfall has greater impact on the occurrence of debris flow than rainfall intensity. This may be expressed by the cumulative characteristics of the rainwater in the soil stratum which is affected by many factors such as land use and soil type. Therefore, to compensate for the cumulative characteristics of rainwater infiltration and occurrence delay, the accumulated effective rainfall (ER) parameter is considered. The parameter ER can be easily calculated by using equation (13)

$$ER = \sum_{t=0}^{p} \alpha^t d_t \tag{13}$$

which presents a decay coefficient α that accounts for the decreasing influence of previous rainfall at time t, which may be calculated by (14). Parameter d is the actual rain (already weighted if necessary, as previously explained), in which $t = 0$, refers to the start of the rainfall period, and p is the actual value for which ER is to be calculated.

$$\alpha = \sqrt{K} \tag{14}$$

DOI: 10.1002/hyp

Where K is a coefficient proposed by Fedora and Beschta (1989) which depends on the sub-basin area A and may be calculated by equation (15):

$$K = 0.881 + 0.00793 \cdot \ln(A) \quad (15)$$

Analysis of morphological factors. Even within the same watershed, different sub-basins will probably present distinct topographic and morphologic characteristics. To improve the accuracy of the proposed prediction model, it is necessary to consider such characteristics and use them as part of the model input information. To do so, a GIS based tool was applied to accurately grasp the main features of the sub-basins, which include the sub-basin area (km^2), length of major rivers of the sub-basin (km), the average slope of the sub-basin (degree) and, finally, a morphological factor called Horton's form factor, F. The form factor parameter is defined as the ratio between the area of the watershed and the square of the total length of the main river.

Application of GIS was found to greatly increase the accuracy and efficiency of morphological data identification and analysis. Table I lists the sites and dates of available data. Table II shows the chosen morphological parameters for representing the sub-basins with their actual obtained values. It is important to note that these values are within the same magnitude. Therefore, if the proposed model is to be applied for the prediction of debris flows for a new sub-basin, such characteristics should be carefully considered, as the proposed model is based on data-driven techniques. Supervised ANN models tend to give more accurate results whenever new input information is within the limits of the training data.

RESULTS AND DISCUSSION

We have already discussed some of the issues related to the destructive power of debris flows and the importance of an early warning system. This section explains the proposed intelligent warning system that can be used for debris flow prediction. Accurate prediction of debris flows is still a very complex and difficult task, particularly regarding the estimation of actual discharge volumes. Hence, in the proposed warning system, only information regarding the 'occurrence' and 'no occurrence' of the debris flows is addressed. The system is based on the previously described SNN + NN model, which combines the clustering capability of the SNN with the learning-from-data ability of the ANN models. In an attempt to define the 'optimal' values of the model's parameters (such as K, KT) and input variables, various experiments were carried out by a trial-and-error procedure, which is described in detail later.

Many other previously published works emphasized that the occurrence of debris flows is not only related to hydrological factors such as rainfall volumes and duration, but also strongly affected by morphological characteristics of the river basin. To find the best architecture (i.e. with the most appropriate input units), four different models have been developed and their efficiency investigated. These models include the two hydrological factors previously described: effective rainfall duration (T) and accumulated effective rainfall (R). However, they vary in the morphological factors used as input information. The four morphologically related factors here are considered for each sub-basin: average slope (S), area (A), length of main rivers and creeks (L) and form factor (F). We can summarize the four models as follows:

- Model 1 : 3-input model–T, R and S
- Model 2 : 4-input model–T, R, S and A
- Model 3 : 5-input model–T, R, S, A and L
- Model 4 : 6-input model–T, R, S, A, L and F

Figure 8 shows a schematic representation of the proposed SNN + NN model including six input variables

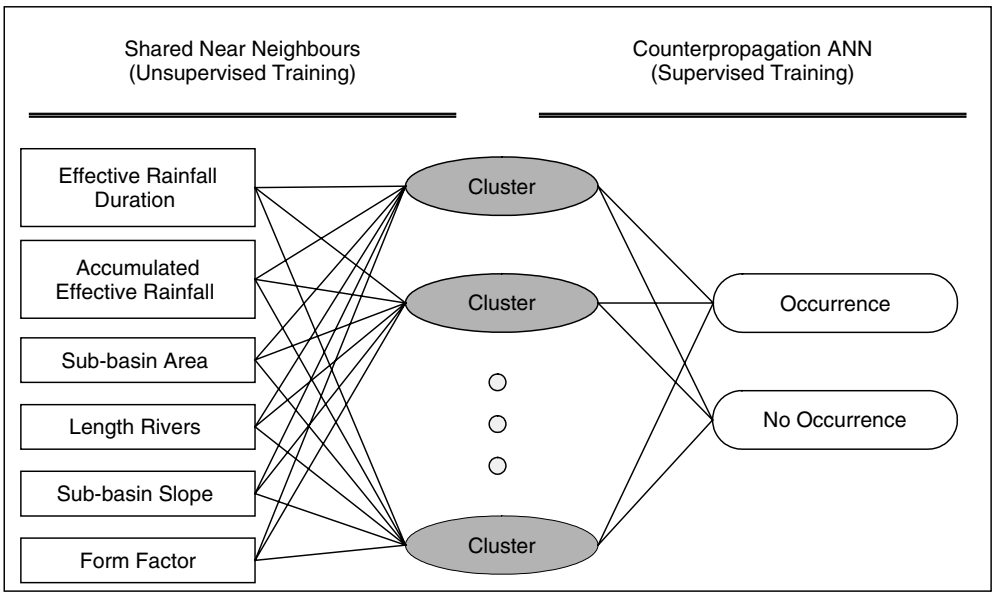

Figure 8. Architecture for one of the proposed SNN + NN models for debris flow warning systems

DOI: 10.1002/hyp

Table III. Comparison of results of the four basic models

Models	K (KT)	No. of clusters	Training set			Testing set		
			wrong prediction / total train data	(Error Ratio)	right prediction / occurred flow	wrong prediction / total test data	(Error Ratio)	right prediction / occurred flow
3-input model (T, R, S)	5 (2)	132	16/620	(2.6%)	12/19	10/418	(2.4%)	3/7
4-input model (T, R, S, A)	7 (4)	139	14/620	(2.3%)	13/19	11/418	(2.6%)	4/7
5-input model (T, R, S, A, L)	5 (2)	112	*15/620*	*(2.4%)*	*13/19*	*10/418*	*(2.4%)*	*5/7*
6-input (T, R, S, A, L, F)	5 (2)	107	16/620	(2.6%)	12/19	9/418	(2.2%)	5/7

Note: gray shaded cells contain the model with the best performance and its results.

(Model 4), and two output units representing the 'occurrence' and 'no occurrence' of debris flow (which is the same for all four models) in the next step (hour). The left part of the network represents the SNN component, which is responsible for the clustering of the input data. And the second part represents the NN component, which is intended to improve the model's final output through a weighted method of hidden layer clusters after a supervised learning process. Note that the number of clusters depends on the values of parameters K and KT of the SNN component.

To evaluate the performance of the four models, the total data set (1038 points) has been divided into two groups, i.e. training and testing periods. The training data set corresponds to 620 points, presenting 19 'occurrences' and 601 'no occurrences' of actual debris flows. For the 418 points used for the testing data, there were seven 'occurrences' and 411 'no occurrences' of actual debris flows. As shown, there are only a few of occurrences. This is mainly because in a specific area we could only obtain at most one point in a heavy rainfall event, which belongs to 'occurrence'.

After specifying the architecture of the four models, we conducted an extensive investigation of different values for K and KT parameters in an attempt to define their optimal values. For each model, values of K between 3 and 9, and KT between 1 and 8 have been tested and their efficiency compared. Once the parameters K and KT were defined and, consequently, the total number of clusters identified, we can then compare the four models (with their optimal values for K and KT) of which the results are also summarized in Table III. In general, all four models performed very well during the training and testing periods, presenting low values for the error ratios of around 2.5 and 2.4%, respectively. It appears that the proposed models can make accurate forecasts of debris flows. Note also, the consistent performances in both training and testing periods present clear evidence of the proposed models being suitable and robust.

Out of the four models investigated, Model 3 (5-input model–T, R, S, A and L) presents the most consistent and good performance, having parameters $K = 5$ and $KT = 2$. Note that the introduction of the six factors (Model 4) does not bring any improvement of results during the training period. However, when applying the model with the testing data set, Model 4 is the one that shows the best performance. According to this analysis, we find the hydrological factors (rainfall volumes and duration) are the most important factors, which dominate whether the debris flow will occur or not, while the morphological factors implemented in this study could delicately provide extra information to judge the system behaviour.

Tables IV and V show more detailed results of Model 3 and its efficiency for the training and the testing periods, respectively. For example, for the testing period (Table V), the developed warning system could correctly predict the 'no occurrence' of debris flows for 403 times out of the 411 (= 403 + 8) times the flow did not actually occur ('no occurrence') resulting in an efficiency ratio of 98.1%. Similarly, it also could predict correctly 5 out of the 7 (= 2 + 5) times the debris flows were actually observed ('occurrence'), yielding an efficiency ratio of 71.4%. The same can be said of the results for the training period shown in Table IV. These results strongly indicate

Table IV. Results of the 5-input SNN model–training period

		Calculated values		Accuracy(%)
		No flow	Flow	
Observed data	No Flow	592	9	98.5
	Flow	6	13	68.4
Overall efficient ratio = 605/620 = 97.6%				

Note: gray shaded cells contain the numbers of right predictions.

Table V. Results of the 5-input SNN model - validation period

		Calculated values		Accuracy(%)
		No flow	Flow	
Observed data	No flow	403	8	98.1
	Flow	2	5	71.4
Overall efficient ratio = 408/418 = 97.6%				

Note: gray shaded cells contain the numbers of right predictions.

DOI: 10.1002/hyp

that the proposed methodology has high accuracy and therefore has been successfully applied for the prediction of occurrence (or not) of debris flow in the form of an intelligent early warning system.

CONCLUSIONS

The main objectives of building a debris flows warning system are to avoid loss of human life and decrease the economic losses caused by such flows, as much as possible. Debris flows are influenced by a range of factors including soil type, geomorphologic factors, land use, rainfall intensity and distribution, and topographic characteristics. Hence, the physical phenomena behind the debris flows present great complexity and high nonlinearity, making accurate prediction an almost impossible task by only using traditional techniques. The paper describes the development and application of a novel type of neural network called SNN + NN in which a clustering mechanism is developed based on the SNN method. The important parameters K and KT of SNN are identified through a trial-and-error investigation for defining their optimal values. This method has, as its greatest advantages, a small computational memory and a small number of parameters (only two, K and KT), which results in a much faster model when compared to the traditional CPN model.

The model input information includes two groups of data: hydrological and morphological. The hydrological data include effective rainfall and rainfall duration, which can incorporate the influences on debris flow related to the time delay of the rainfall event. Moreover, it was also shown that consideration of morphological data as input information, such as slope, length of river, basin area and form factor, increased the model's performance. The model performance was evaluated based on the efficiency ratio. For debris flow warning of 'no occurrence', the model was 98% accurate for both training and testing periods. In the case of 'occurrence', even though poorer than the previous results, the model could still perform very well presenting right predictions of around 70% for the simulated values. Therefore, the proposed methodology has proved to be able to successfully predict the occurrence, or not, of debris flows in the form of an intelligent early warning system.

ACKNOWLEDGEMENTS

The authors offer their deepest thanks for the funds received during this research from the Water and Soil Conservation Bureau, ROC [SWCB-92-012-10]. Moreover, the authors are indebted to professor Zheng-Cheng Fan, National Taiwan University, for his constructive suggestions and for providing valuable data for our research.

REFERENCES

Ayalew L, Yamagishi H. 2005. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan. *Geomorphology* **65**(1–2): 15–31.

Casadel M, Dietrich WE, Miller NL. 2003. Testing a model for predicting the timing and location of shallow landslide in soil-mantled landscapes. *Earth Surface Processes and Landforms* **28**: 925–950.

Chang FJ, Hwang YY. 1999. A self-organization algorithm for real-time flood forecast. *Journal of Hydrology* **169**: 229–241.

Chang LC, Chang FJ. 2001. Intelligent control for modeling of real time reservoir operation. *Hydrological Processes* **15**: 1621–1634.

Chang FJ, Hu HF, Chen YC. Counterpropagation fuzzy-neural network for stream flow reconstructing. *Hydrological Processes* 2001. **15**(2): 219–232.

Chang LC, Chang FJ, Tsai YH. The fuzzy exemplar-based inference system for flood forecasting. *Water Resources Research* 2005. **41**: W02005.

Chau KT, Sze YL, Fung MK, Wong WY, Fong EL, Chan LCP. 2004. Landslide hazard analysis for Hong Kong using landslide inventory and GIS. *Computers & Geosciences* **30**(4): 429–443.

Crozier MJ. 1999. Prediction of rainfall-triggered landslides: a test of the antecedent water status model. *Earth Surface Processes and Landforms* **24**: 825–833.

Ermini L, Catani F, Casagli N. 2005. Artificial Neural Networks applied to landslide susceptibility assessment. *Geomorphology* **66**(1–4): 327–343.

Fedora MA, Beschta RL. 1989. Peak flow simulation using an antecedent precipitation index (API) model. *Journal of Hydrology* **112**: 121–133.

Gómez H, Kavzoglu T. 2005. Assessment of shallow landslide susceptibility using artificial neural networks in Jabonosa River Basin, Venezuela. *Engineering Geology* **78**(1–2): 11–27.

Hecht-Nielsen R. 1987. Counterpropagation network. *Applied Optics* **26**: 4979–4984.

Jarvis RA, Patrick EA. 1973. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers* **C**(22): 1025–1034.

Lan HX, Zhou CH, Wang LJ, Zhang HY, Li RH. 2004. Landslide hazard spatial analysis and prediction using GIS in the Xiaojiang watershed, Yunnan, China. *Engineering Geology* **76**(1–2): 109–128.

Lee S, Ryu JH, Won JS, Park HJ. 2004. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. *Engineering Geology* **71**(3–4): 289–302.

Lollino G, Arattano M, Cuccureddu M. 2002. The use of the automatic inclinometric system for landslide early warning: the case of Cabella Ligure (North-Western Italy). *Physics and Chemistry of the Earth* **27**: 1545–1550.

Perotto-Baldiviezo HL, Thurow TL, Smith CT, Fisher RF, Wu XB. 2004. GIS-based spatial analysis and modeling for landslide hazard assessment in steep lands, southern Honduras. *Agriculture Ecosystems & Environment* **103**(1): 165–176.

Sajikumar N, Thandaveswara BS. 1999. A non-linear rainfall–runoff model using an artificial network. *Journal of Hydrology* **216**: 32–55.

Shamseldin AY. 1997. Application of a neural network technique to rainfall–runoff modelling. *Journal of Hydrology* **199**: 272–294.

Yang HC, Chang FJ. 2005. Modelling the combined open channel flow by artificial neural network. *Hydrological Processes* **19**: 3747–3762.

Yesilnacar E, Topal T. 2005. Landslide susceptibility mapping: a comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). *Engineering Geology* **79**(3–4): 251–266.

Yu FC. 2002. An overview of debris flow in Taiwan. In *First International Conference on Debris-flow Disaster Mitigation Strategy*, Taipei, Taiwan, 185–196.

DOI: 10.1002/hyp